Intrinsic Judgment Error in Men's Championship World Surf League: WSL 2021

Supplementary materials:
https://osf.io/yk2vt/

For correspondence:
tony.meireles@ufpe.br

Tony Meireles Santos[1,2], Lucas Eduardo Rodrigues Santos[2], Ítalo Vinícius[3], Cayque Brietzke[4], Lucas Camilo Pereira[5], Paulo Henrique Melo[1], Thaiene Camila Beltrão Moura[1], Hassan Mohamed Elsangedy[5], Flávio Oliveira Pires[4]

[1]Grad Program in Physical Education, Federal University of Pernambuco, Pernambuco, Brazil. [2]Grad Program in Neuropsychiatry and Behavior Sciences, Federal University of Pernambuco, Pernambuco, Brazil. [3]Grad Program in Physical Activity Sciences, School of Arts, Sciences and Humanities at the University of São Paulo, São Paulo, Brazil. [4]Grad Program in Human Movement Science and Rehabilitation Program, Federal University of São Paulo, São Paulo, Brazil. [5]Grad Program in Physical Education, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil.

Please cite as: T. M. Santos, L. E. R. Santos, Í. Vinícius, *et al.* (2022). Intrinsic Judgment Error in Men's Championship World Surf League: WSL 2021. *SportRχiv.*

All authors have read and approved this version of the manuscript. This article was last modified on July 25, 2022.

Authors TMS @tonymsantos and LERS @LucasERSantos can be reached on Twitter.

ABSTRACT

Surfers' performances are subjectively ranked by 5 judges. Low reliability and validity in judgment may lead to preventable errors and unfair scores. Aiming to describe the judgment error we analyzed the available WSL's data related to 2021 Men's Championship Tour (4,095 waves; 20,475 scores). We found an inverted 'U'-shape pattern for the judgment error vs. control score, explained by a quadratic regression model (R = 0.52; SEE = 0.10). The reliability produced excellent Intraclass Correlation Coeficient ($CI_{95\%}$ = 0.97, 1.00), with a between judge (typical) error of 0.15. Validity analyses indicated a minimal real difference of 0.49 in the sum of two waves between the surfers for having 95% certainty for the heat winner. We recommend WSL to incorporate the intrinsic judgment error in into judgments for increasing the fairness and trust on WSL championship tour.

INTRODUCTION

Competitive sports are dependent on refereeing for conduct the events and reporting infractions. In several modalities, such as surfing, the panel of judges assumes responsibility for defining the results of disputes through the attribution of scores for each athletic performance (surfed wave). The judging criteria consider the commitment and the degree of difficulty of the surfed wave, as well as the performance of innovative, progressive, and combined maneuvers, varying the speed, power and flow in the execution [1]. The evaluations even starting from pre-defined criteria and a rigorous training process carry some degree of subjectivity, increasing judgment variability that technically can be characterized as an error, eventually culminating in dubious results. According to Kahneman and Frederick [2], human judgment occurs through an interaction between different cognitive systems, in which judges make their assessments through conscious and unconscious thoughts, which can sometimes hide implicit biases even without intention [3].

The sports in which athletes' performance are determined by scores (i.e. gymnastics, skating, etc.), traditionally try to explain atypical results by the existence of different types of intentional bias (i.e. nationality, sequential and conformity) [4]. However, error in judgment is a natural and inevitable phenomenon, making it necessary to know and manage them. Even with all care taken to properly conduct trials, some judges are more accurate than others [5]. This fact can occur due to different experience times in the modality or in sensitivity to interpret the surfer performance. In addition to the inherent flaws in the entire judging process [6, 7], surf competitions are exposed to wave quality predictions, causing eventual delays or postponements of the start of the competition. These competitive daily routines impose on judges long working hours (< 10 h), in which even considering the existing rotation, results in deprivation of rest, inadequate nutrition and long moments of sustained attention, which can cause mental fatigue [8].

Several other combined elements can increase the complexity of judgments. Compared to other modalities, surf presents greater intervention of random variables that generate increases of unpredictability in the performance of surfers, and possibly in the evaluation of the judges. The competition is performed in a wide area to be observed, with different visual perspectives and time intervals between each wave. The waves often have huge characteristics variability, since changes in direction, shape and size, sea conditions, winds and weather, which influence the maneuvers performed by surfers. In addition, according to the rule practiced by the World Surf League (WSL) in 2021, there is no limit to the number of waves surfed in a heat. This makes a need to publish the scores almost immediately, giving the judges a sense of urgency in the definition of their multiple judgments in a short period of time. Those aspects could compromise the clinimetrics properties of judgment in surf.

The low reliability and validity of judgments can imply in poorly reproducible results that provoke in inaccuracies and injustices. In surfing, reliable results should show how much a judge's scores are internally consistent with their peers (between judges), indicating more uniformity in the evaluation of performance. Judgment validity is when the scores assigned by a judge accurately reflect the surfer's final score (gold standard). Some studies have investigated the judge’s performance using reliability and validity metrics to determine competence indexes on judgment [9, 10]. However, it is possible that such indexes offer low practical applicability for improve quality and justice on the results of competitions.

The reliability metrics used in the literature usually determine the relationship between the judgment scores, either utilizing the intraclass correlation coefficient (ICC) [9-11] or non-parametric statistics like Cochran's Q [9]. To complement these measures, it is suggested the use of statistics such as the standard error of measurement (SEM), which reflects the measurement error in an absolute way, as reported by Premelč, et al. [11]. On the other hand, studies that claim to determine the validity of judgment in different sports [10, 12], need more rigor in detailing the procedures regarding the validity model used. A common limitation of these studies is the use of metrics such as ANOVA's, Kendall's W coefficient and Theta coefficient, which do not provide the dimensioning of error in judgment. For that, the use of the Intrinsic Judging Error Variability (IJEV) [4, 13] has been recently proposed, and similarly to the typical measurement error [14], offers an approximation to dimensioning of error.

In this sense, it is necessary to know the error of judgment in surf to raise evidence and discussions about its management. The way in which sports entities (e.g. WSL) deal with the error of judgment is still not well understood, and this can negatively impact the surfing community. For example, in the case of athletes, errors in judgment can harm sports careers, with direct (award) or indirect (sponsorship, campaigns, etc.) financial implications, staying in the world surfing elite, mental health problems, among others. In those context, the present article has three aims. First, it was to establish in an exploratory character the behavior of the error of judgment of each wave surfed according to its assigned score, as proposed by Heiniger and Mercier [4]. Second, was to establish the between-judge reliability of the 2021 WSL judges' panel. Finally, and probably the most relevant implications for the modality, it was to establish the judgment validity by the comparison of each judge score with the final control score of the wave (CS). Considering the clinimetric aspect of the paper, that isn't an a priori hypothesis.

METHOD

Database

Methodological study based on database analysis to determine the reliability (relative and absolute) and validity of the WSL's judges. Available data from all waves surfed (n = 4095) by male participants of the elite world surf (n = 55) in 7 events in 2021 tour were analyzed (data from Jeep Surf Ranch Pro, was not available). This study was conducted by researchers with any kind of relationship with the athletes, judges or WSL. The overview of selection and data treatment steps is presented in Figure 1.

Procedures

Data extraction was performed manually by the study authors directly from the WSL website (public access - https://www.worldsurfleague.com) between September and December/ 2021. Data were extracted by pairs of independent researchers randomly assigned, in which each researcher was responsible for extracting at most two events. The scores of each judge for wave surfed were recorded in a specific form and later included in a spreadsheet. After merging the two worksheets at each event, typos and miscellaneous inconsistencies were registered for later confirmation (Figure 1). The validation of the extractions performed was carried out by a different pair of researchers. Then, the combined spreadsheets returned to the pair of researchers originally responsible for extracting the data to work together to fix the identified inconsistencies. After a new verification and resolution of all problems, the data from a given moment were then integrated into the general database with all events, which was later used for analysis of present study.

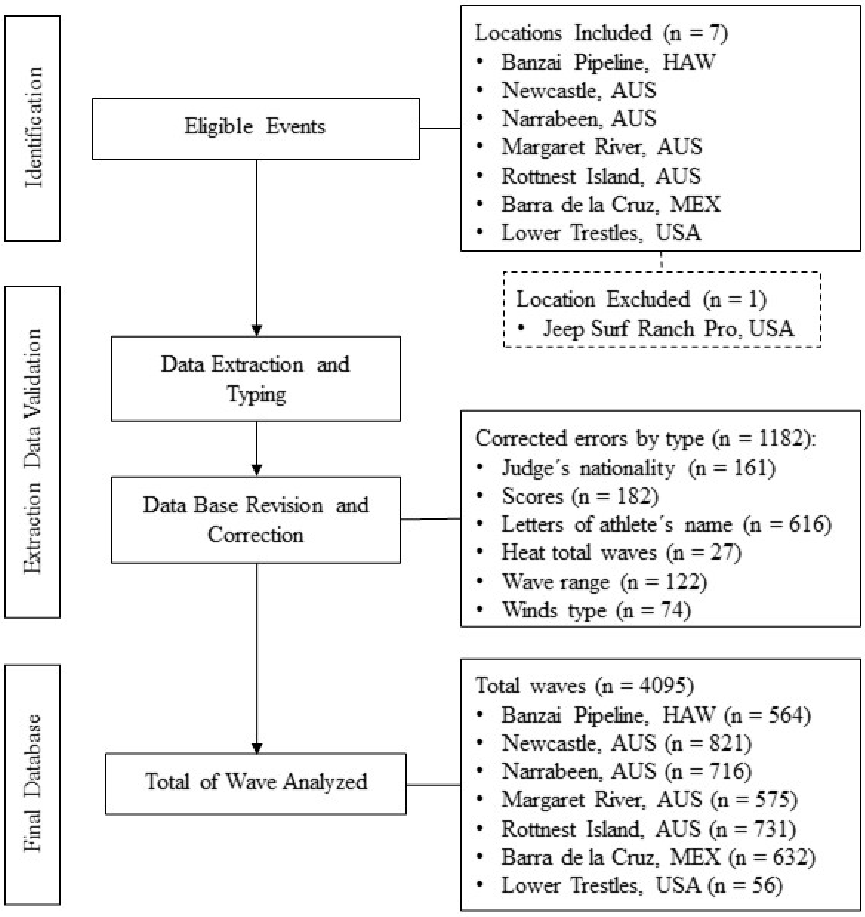
Identification
Extraction Data Validation
Final Database
Eligible Events
Locations Included (n = 7)
• Banzai Pipeline, HAW
• Newcastle, AUS
• Narrabeen, AUS
• Margaret River, AUS
• Rottnest Island, AUS
• Barra de la Cruz, MEX
• Lower Trestles, USA
Location Excluded (n = 1)
• Jeep Surf Ranch Pro, USA
Data Extraction and Typing
Data Base Revision and Correction
Corrected errors by type (n = 1182):
• Judge´s nationality (n = 161)
• Scores (n = 182)
• Letters of athlete´s name (n = 616)
• Heat total waves (n = 27)
• Wave range (n = 122)
• Winds type (n = 74)
Total of Wave Analyzed
Total waves (n = 4095)
• Banzai Pipeline, HAW (n = 564)
• Newcastle, AUS (n = 821)
• Narrabeen, AUS (n = 716)
• Margaret River, AUS (n = 575)
• Rottnest Island, AUS (n = 731)
• Barra de la Cruz, MEX (n = 632)
• Lower Trestles, USA (n = 56)

Figure 1 - Flow chart of data extraction and analysis

Source: Authors

Statistical Analysis

For the overview of the judgment error, the Intrinsic Judging Error Variability (IJEV) was adopted as proposed by Heiniger and Mercier [13] (Equation 1). The exploratory characteristics of this variable enable the visualization of the error behavior of each wave judgment relative to its respective CS. After an exploratory approach, the best model for a regression analysis was presented between IJEV (dependent variable) vs. CS (predictor variable). As result, the determination coefficient (R), p-value, standard estimation error (SEE) and the prediction equation were presented.

$$IJEV = SD\ [J_{Diff\_1}, J_{Diff\_2}, J_{Diff\_3}, J_{Diff\_4}, J_{Diff\_5}] \qquad \text{Eq. 1}$$

Where:

IJEV - Intrinsic judging error variability

SD - standard deviation

J_{Diff_1} - Difference between the control score and the judge score for judge 1 (J_{Diff_2} for judge 2 and so on)

The intraclass correlation coefficient was used considering the model of one-way random effects, mean of k raters (n = 5) and absolute agreement ($ICC_{(1,k)}$), as recommended by Koo and Li [15]. The confidence interval for 95% ($CI_{95\%}$) and the level of significance to explore the relative reliability between judges (inter-judge reliability) were also calculated. The ICC was classified as follows: < 0.5 (Poor); between 0.5 and 0.75 (Moderate); between 0.75 and 0.9 (Good); and > 0.9 (Excellent), as suggested by the same authors. For absolute reliability, the standard error between judges (SEB_J) was used for estimating the error of judge compared to his peers (Equation 2). Besides, the minimal real difference of judges (MRD_J) was calculated to representing the threshold of a real error (Equation 3) between judges.

$$SEB_J = SD \times (\sqrt{}1 - ICC) \qquad \text{Eq. 2}$$

Where:

SEB_J - Standard error between judges

SD - standard deviation

ICC - Intraclass correlation coefficient

$$MRD_J = SEb_J \times 1{,}96 \times \sqrt{2} \qquad \text{Eq. 3}$$

Where:

MRD_J - Minimal real difference between judges

SEB_J - Standard error between judges

Considering that the reliability promotes a between judge analysis, we also explored the impact of the error magnitude on the difference between each judgment with the CS - previously described as a 'special case of validity' [9]. Due to the absence of a 'gold standard' method for wave assessment, the available option is the utilization of a central tendency parameter of the panel of judges. For this purpose, the median of five judges was used as the CS based on the following arguments presented by Heiniger and Mercier [4]: a. to be the best proxy of an 'actual wave score'; and b. could be more robust against misjudgments and biased judges compared to the trimmed average (approach utilized by WSL). Moreover, could be keep in mind other arguments to use the median: c. the low number of wave scores (n = 5) to generate a normal distribution enabling an average calculation; and d. your robustness to prevent interferences from an erratic score (discrepant or outlier). The magnitude of the error for validity for one wave was established by the typical error of judgement in reference to CS (TEJ_{CS_1W}, Equation 4a) and its superior confidence interval for 95% defined as minimal real difference for CS (MRD_{CS_1W}, Equation 5a), a new variable mixing a traditional approach for error measurement in sports science (typical error of measurement equal to the standard deviation of the differences divided by square root) [14]. Because WSL utilize the sum of two waves in comparison of athletes in the heats, TEJ_{CS_1W} and MRD_{CS_1W} were also presented for the sum of two waves by multiplication by 2 (TEJ_{CS_2W} and MRD_{CS_2W}, respectively).

(a) $TEJ_{CS\_1W} = SD_{Overall}\ [J_{Diff\_1}, J_{Diff\_2}, J_{Diff\_3}, J_{Diff\_4}, J_{Diff\_5}] \div \sqrt{2}$ Eq. 4

(b) $TEJ_{CS\_2W} = TEJ_{CS\_1W} \times 2$

Where:

TEJ_{CS_1W} - Typical error of judgement for control score for one wave

TEJ_{CS_2W} - Typical error of judgement for control score for two waves

$SD_{Overall}$ - Standard deviation for the data matrix

J_{Diff_1} - Difference between the control score and the judge score for judge 1 (J_{Diff_2} for judge 2 and so on)

$MRD_{CS\_1W} = \sqrt{}\ (DF \times TEJ_{CS\_1W}^2\ ) \div \chi^2_{97.5\%}$ Eq. 5

$MRD_{CS\_2W} = MRD_{CS\_1W} \times 2$

Where:

MRD_{CS_1W} - Minimal real difference for control score for one wave

MRD_{CS_2W} - Minimal real difference for control score for two waves

DF - Degrees of freedom

TEJ_{CS_1W} - Typical error of judgement for control score

$\chi 2$ - Chi square

Analyzes of reliability and validity were performed for the overall database and for the subgroups of interest: location, round in regular competition, round in finals, wave level, number of athletes in the heat, wave size and wind condition. The dataset was organized in Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) and later analyzed at Rstudio [16]. Intraclass correlation were calculated through Psych package [17], while IJEV, SEB_J, MRD_J, TEJ_{CS_1W}, and MRD_{CS_1W} were calculated by using Rstudio native functions. The level of significance was adjusted for $p < 0.05$.

Results

The representation of the IJEV for each surfed wave in WSL 2021 championship as a function of the CS, is presented in figure 2. The inverted 'U' pattern described by the second order (quadratic) polynomial model produced an R = 0.52, $p < 0.001$ and SEE = 0.1038. The smallest magnitudes of IJEV were observed for waves classified as Poor ($\cong 0.18$) and Excellent ($\cong 0.28$), and higher for waves classified as Good ($\cong 0.40$).

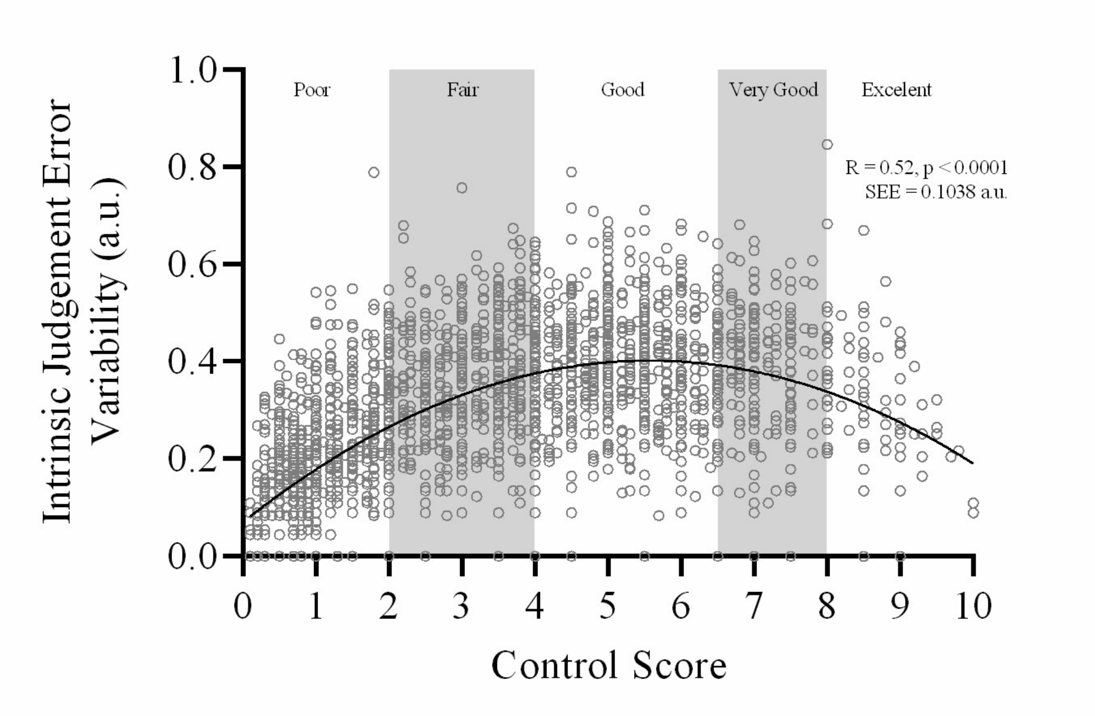

Figure 2 - Intrinsic Judging Error Variability by Control Score for all 2021 season. R - coefficient of determination; SEE - standard estimation error; a.u. - arbitrary unit; the vertical tracks represent 2021 World Surf League criteria for classification the wave quality; The prediction equation for IJEV based on second order polynomial quadratic model is IJEV = 0,06966 + 0,1192 x CS - 0,01070 x CS^2

Source: Authors

The global reliability of judgments performed in 2021 (overall) or segmented by the conditions of interest (location, rounds, wave level, number of athletes in the water, wave size and wind intensity) were reported in Table 1. The relative reliability, expressed by ICC, confidence interval and p value suggest that the WSL judges showed 'near perfect' performance, with ICC indices consistently close to 1 ($CI_{95\%}$ = 0.970, 0.996; p < 0,001) and classified in most times (92%) as Excellent. Good classification (8%) was observed only when data were segmented by wave level. The absolute reliability presented by between judge error was stable for the average of the whole dataset results, with $SEB_J \cong 0.15$ (5.3%) and $MRD_J \cong$ 0.41 (14.6%).

Table 1 - Overall and categorized relative and absolute reliability for surf judgments during 2021 WSL surf season for males

Variables of Interest	n	Relative				Absolute	
		$ICC_{(1,k)}$	Class.	$CI_{95\%}$	p value	SEb_J	MRD_J
						Raw	Raw
Overall	4095	0,996	E	0.996, 0.996	< 0.001	0,14	0,40
By Location							
Banzai Pipeline, HAW	564	0,997	E	0.996, 0.997	< 0.001	0,14	0,38
Newcastle, AUS	821	0,996	E	0.995, 0.996	< 0.001	0,15	0,40
Narrabeen, AUS	716	0,996	E	0.995, 0.996	< 0.001	0,14	0,38
Margaret River, AUS	575	0,996	E	0.996, 0.997	< 0.001	0,15	0,42
Rottnest Island, AUS	731	0,997	E	0.996, 0.997	< 0.001	0,14	0,39
Barra de la Cruz, MEX	632	0,996	E	0.996, 0.997	< 0.001	0,15	0,41
Lower Trestles, EUA	56	0,998	E	0.998, 0.999	< 0.001	0,13	0,37
By Round in Regular Competition							
Seeding Round	1207	0,996	E	0.996, 0.996	< 0.001	0,15	0,41
Elimination Round	350	0,996	E	0.995, 0.996	< 0.001	0,15	0,41
Round 32	1271	0,996	E	0.996, 0.997	< 0.001	0,15	0,40
Round 16	678	0,997	E	0.996, 0.997	< 0.001	0,14	0,38
Quarter Final	290	0,997	E	0.997, 0.998	< 0.001	0,14	0,39
Semin Final	154	0,997	E	0.997, 0.998	< 0.001	0,14	0,38
Finals	89	0,998	E	0.997, 0.999	< 0.001	0,13	0,37
By Round in Finals							
Match 1	10	0,995	E	0.990, 0.998	< 0.001	0,16	0,44
Match 2	7	0,998	E	0.994, 0.999	< 0.001	0,14	0,38
Match 3	8	0,999	E	0.997, 1.000	< 0.001	0,11	0,30
Match 4	31	0,999	E	0.998, 0.999	< 0.001	0,12	0,33
By Wave Level							
Poor	1605	0,968	E	0.966, 0.970	< 0.001	0,09	0,24
Fair	1297	0,963	E	0.960, 0.966	< 0.001	0,18	0,49
Good	659	0,828	G	0.810, 0.845	< 0.001	0,23	0,65
Very Good	398	0,827	G	0.803, 0.849	< 0.001	0,23	0,63

Excellent	136	0,898	G	0.873, 0.919	< 0.001	0,19	0,53
By Number of athletes in the heat							
Two	2538	0,996	E	0.996, 0.997	< 0.001	0,14	0,40
Three	1557	0,995	E	0.994, 0.995	< 0.001	0,15	0,42
By Wave Size							
1 to 4	2629	0,996	E	0.996, 0.996	< 0.001	0,14	0,40
4 to 6	766	0,997	E	0.997, 0.998	< 0.001	0,14	0,39
6 to 8	298	0,996	E	0.996, 0.997	< 0.001	0,15	0,42
8 to 10	121	0,997	E	0.996, 0.997	< 0.001	0,15	0,41
Not Reported	281	0,997	E	0.996, 0.997	< 0.001	0,15	0,41
By Wind Conditions							
Calm	1768	0,996	E	0.996, 0.997	< 0.001	0,14	0,40
Cross	329	0,995	E	0.994, 0.995	< 0.001	0,15	0,42
Light	536	0,997	E	0.996, 0.997	< 0.001	0,14	0,39
Offshore	1167	0,997	E	0.996, 0.997	< 0.001	0,14	0,39
Onshore	14	0,990	E	0.982, 0.996	< 0.001	0,14	0,39
Not Reported	281	0,997	E	0.996, 0.997	< 0.001	0,15	0,41

Legend: n - Number of waves; $ICC_{(1,K)}$ - Intraclass correlation model one way random (1,5); Class. - ICC classification; $CI_{95\%}$ - ICC confidence interval for 95%; p value - ICC p value; SEBJ - Standard error between judges; MRDJ - Minimal real difference between Judges.

Source: Authors

The judgment errors to determine the surfer final score (e.g. CS) considering different variables of interest are presented in Table 2. For all waves surfed (Overall), a result of 0.22 and 0.25 was found for TEJ_{CS_1W} and MRD_{CS_1W}, respectively. As the WSL uses the sum of two best waves as a criterion to compare the performance between surfers, the results of TEJ_{CS_2W} (0.44) and MRD_{CS_2W} (0.49) were available. Most of the investigated conditions with potential for judgment disruption, resulted in a low mean variability for TEJ_{CS_1w} ($\cong$ 0.22) of the whole dataset. Considering the segmented conditions, a low mean variation of TEJ_{CS_1w} was also found for different locations that hosted the events ($CI_{95\%}$ = 0.21, 0.23), different rounds of regular competition ($CI_{95\%}$ = 0.21, 0.22), wave size ($CI_{95\%}$ = 0.21, 0.23) and wind condition ($CI_{95\%}$ = 0.21, 0.23). On the other hand, a slightly higher mean variation was found for the analyzes per round in the finals ($CI_{95\%}$ = 0.15, 0.29), by wave level ($CI_{95\%}$ = 0.16, 0.31) and by number of surfers disputing the heat ($CI_{95\%}$ = 0.17, 0.26).

Table 2. Overall and categorized validity for surf judgments during 2021 WSL surf season for males

Variables of Interest	One Wave		Sum of Two Waves	
	TEJ_{CS_1W} (a.u.)	MRD_{CS} (a.u.)	TEJ_{CS} (a.u.)	MRD_{CS} (a.u.)
Overall	0.22	0.25	0.44	0.49
By Location				
Banzai Pipeline. HAW	0.21	0.24	0.42	0.48
Newcastle. AUS	0.22	0.24	0.44	0.48
Narrabeen. AUS	0.21	0.24	0.42	0.48
Margaret River. AUS	0.23	0.24	0.47	0.48
Rottnest Island. AUS	0.22	0.24	0.43	0.48
Barra de la Cruz. MEX	0.23	0.24	0.46	0.48
Lower Trestles. EUA	0.21	0.22	0.42	0.43
By Round in Regular Competition				
Seeding Round	0.22	0.24	0.45	0.48
Elimination Round	0.23	0.24	0.45	0.47
Round 32	0.22	0.24	0.44	0.48
Round 16	0.21	0.24	0.42	0.48
Quarter Final	0.21	0.23	0.42	0.47
Semin Final	0.21	0.23	0.42	0.46
Final	0.21	0.22	0.42	0.43
By Round in Finals				
Match 1	0.28	0.18	0.55	0.36
Match 2	0.23	0.17	0.45	0.35
Match 3	0.19	0.18	0.38	0.35
Match 4	0.18	0.21	0.36	0.41
By Wave Level				

Poor (< 2.0)	0.12	0.24	0.25	0.49
Fair (2.0 - 3.9)	0.25	0.24	0.51	0.48
Good (4.0 - 6.4)	0.28	0.24	0.55	0.48
Very Good (6.5 - 7.9)	0.27	0.24	0.54	0.47
Excellent (≥ 8)	0.25	0.23	0.49	0.45
By Number of Athletes in the Heat				
Two	0.22	0.24	0.43	0.49
Three	0.22	0.24	0.45	0.49
By Wave Size				
1 to 4	0.22	0.24	0.44	0.49
4 to 6	0.21	0.24	0.43	0.48
6 to 8	0.23	0.23	0.46	0.47
8 to 10	0.23	0.23	0.45	0.45
Not Reported	0.23	0.23	0.45	0.47
By Wind Conditions				
Calm	0.22	0.24	0.44	0.49
Cross	0.23	0.23	0.47	0.47
Light	0.21	0.24	0.43	0.48
Offshore	0.22	0.24	0.43	0.48
Onshore	0.22	0.19	0.44	0.38
Not Reported	0.23	0.23	0.45	0.47

Legend: TEJ$_{CS\_1W}$ - Typical error of judgement for control score for one wave; TEJ$_{CS\_2W}$ - for two waves; MRD$_{CS\_1W}$ - Minimal real difference for judgement for control score for one wave; MRD$_{CS\_2W}$ - for two waves.

Source: Authors

Discussion

The present study was the first to explore intrinsic judgment error in male professional surf championships organized by the WSL, analyzing the IJEV results, reliability and validity. The judgment error was described for global scores and different conditions of interest (e.g. location, round, wave size, number of athletes inv the heat, wave level and wind). In addition, the use of TEJ_{CS} and MRD_{CS} was proposed to compare the performance of surfers similarly to has been practiced in the interpretation of statistical tests in clinical areas of health and sports performance, incorporating the measurement error for its prognostic relevance [18]. As far as could be observed, there are no studies with other sport modality dedicated to establishing the magnitude of the judgment error and, mainly, its incorporation in the interpretation of the competitive results.

Intrinsic Judgment Error Variability

The behavior of IJEV as a function of CS depends on the modality investigated. Heiniger and Mercier [4] described three main possible kinetics: a) Descending (snowboard halfpipe, acrobatic gymnastics, aerobic gymnastics, artistic gymnastics, rhythmic gymnastics, and artistic swimming); b) 'U' pattern (standard and artistic presentation on dressage) and, as in the present study; c) Inverted 'U' pattern (diving, figure skating, ski jumping, snowboard slopestyle, trampoline and aerials from skiing). According to the authors, the different kinetic patterns are influenced by number of items to be evaluated in each modality, as well as number of errors to be deducted from a given execution.

In addition, the shape of parabola seems to depend on judgments with results close to zero [4], which occurs more often in surfing (40.07% of waves surfed in 2021 season were classified as Poor, 32.0% as Fair, 16.1% as Good, 8.9% as Very Good, and 2.9% as Excellent). It is possible that the low IJEV observed in waves classified as Poor (wave score < 2) is determined by the lower complexity of judgment considering the low number of elements to be observed, since these scores are usually attributed to a wave with incomplete maneuver or surfer mistake. At the opposite end of scale in waves with higher scores, the smallest error observed may be related to a ceiling effect of judgment process provided by proximity of perfection of evaluated performances, which may facilitate the process. Jointly, these results partially confirm the arguments of [4] who suggest that "judges are more accurate when evaluating outstanding or atrocious performances than when evaluating mediocre ones". Based on present results, the WSL judges achieve the better consistency evaluating Poor waves.

The magnitude of predictive power observed on weighted least-squares exponential regression models developed for other sports modalities resulted in superior values of R (0.75 ± 0.19, $CI_{95\%}$ = 0.65, 0.84) and inferior values for root mean square standard deviation (0.14 ± 0.28, $CI_{95\%}$ = 0.00, 0.29) compared to the present study (0.52 and 0.10, respectively). However, those differences appear to be determined by a different process to generate regression models between the present study and the aforementioned paper. While in the present study the full database was utilized to generate our regression, [4] used the mean value for each CS as predictable variable, resulting in a substantial reduction in residuals with directly impact in root mean standard deviation (≅ -48%) and inflating the R (≅ +37%). Considering this and the purpose of the present study related to error scaling, this direct comparison between studies could not be possible.

Relative and Absolute Reliability

Reliability, especially relative, is one of most used metrics in the investigation of psychometric quality of judgment in sports. The relative reliability of present study for the overall data base ($ICC_{(1,k)}$ ≅ 0.99, Table 1), with the exception of segmentation by wave level ($ICC_{(1,k)}$ ≅ 0.90), proved to be much higher than the results of Premelč, et al. [11] for different categories of dance sport (≅ 0.62), by Pajek, et al. [10] for artistic gymnastics in different competitive phases and apparatus (≅ 0.83) and Leandro, et al. [9] in rhythmic gymnastics for athletes in different ranking positions (≅ 0.66).

The absolute dimensioning of the inter-judge error was produced only for sports dance [11], a modality with an evaluation scale like surf (0 to 10), but with a competitive dynamic that results in higher mean scores. The SEM reported in dance was 0.54 (overall), 0.56 (technical qualities), 0.67 (movement to music), 0.57 (partnering skills), and 0.54 (choreography and performance). In the present study, low SEB_J and MRD_J were found for the average of all conditions investigated, with mean values of 0.15 and 0.41, respectively. Considering the higher complexity in surfing judgment, the lowest results of the present study, despite the different calculation strategy, were considered unusual, deserving future investigations to identify which procedural elements are practiced by WSL deserving of popularization.

The interpretation of absolute reliability produces a measure of error between the judges' scores when compared with each other (e.g. judge 1 vs. judge 2, Judge 2 vs. Judge 3 etc.), serving to dimension their qualification. For the total number of waves (n = 4095) surfed in 2021 using the SEB_J (0.14) as criterion for the difference between the judges, expressing the mean error, a frequency of 65% of effectively different scores was observed. When using the MRD_J (0.40) which expresses a 95% certainty of different scores, the frequency was 36%. It is observed that, despite the very high relative reliability scores, the analysis of the results using absolute reliability broadens the understanding of differences between the scores given by the judges.

Since reliability analysis does not describe the error of judge's measurement in relation to CS, its practical application becomes limited for evaluating the competitive dynamics of modality using only the values of SEB_J and MRD_J. In addition, the interpretation of the SEB_J should be performed with caution in the event of heteroscedasticity in the scores [19]. Therefore, the present study produced the results of TEJ_{CS_1W} and MRD_{CS_1W}.

Validity

To the best of our knowledge, two studies investigated the validity of judgment comparing the result of judges with an CS, both in artistic gymnastics [10, 12]. In general, unclear methodological details are observed for the adequate understanding of performed analyses or aims intended for specific purposes (i.e. nationality bias, sequential bias and comparisons between equipment in gymnastics). Leskošek, et al. [12] produced validity indexes considered satisfactory by the authors, while Pajek, et al. [10] interpreted results as unsatisfactory. For this, the authors used ANOVA and Kendall W analysis, which makes it impossible to compare with the findings produced in present study, because those techniques do not measure the error.

Based on the results produced in present study, a concern spot was identified. To define the winner of a heat, the WSL currently uses a difference of 0.01 between the sum of the best two waves surfed, which is below what is necessary to contemplate the natural error of its judging process. The TEJ_{CS} and MRD_{CS} results provide an overview of magnitude of difference required between two surfers for a winner to be defined with a low probability (< 5%) of a random and possibly wrong and unfair result. An applied description of this concept utilization is available in Figure 3 as supplementary material (https://osf.io/yk2vt/) exploring final of the event MEO Pro (Peniche, PT) held in March 2022. In this heat, the difference of the winner (Griffin Colapinto, USA) to the second place (Filipe Toledo, BRA) was 0.14, lower than TEJ_{CS_2W}.

The results produced in present study can be extended to all judging sport modalities, with error magnitudes that need to be established. The use of judgment error would improve the judgment process, establishing certainty (95%) in the definition of winning and losing athletes, reducing the judgment bias and the improvement of sense of justice resulting in an important advancement by WSL to improve its evaluation routines. Previously, modalities such as gymnastics [4] made changes in judgment to make the process more objective and potentially justice. Utilizing an innovative statistical strategy, the present study dimensioned the error of judgment in WSL in 2021 season.

The results produced in present study, analyzed considering their limitations, can offer new perspectives to surf. The main limitations of our study were the analyzes using only data from the competition held in 2021, restricting a possible evaluation of the performance of judges in different years, thus enabling the production of a historical series. In addition, other population groups (i.e. women, surfers in the access and youth categories) and surfing modalities (i.e. long board, big wave, etc.) need to be investigated. Furthermore, possible reasons for the error were not investigated, as previously done in other sports, such as gymnastics [13] and in ski jumping [20].

Conclusion

In conclusion in an applied perspective, our results showed the dimension (0.22 and 0.25 for TEJ_{CS_1W} and MRD_{CS_1W}, respectively) and the inevitable existence of error in WSL judgment. This implies the need to consider the error inherent in this type of evaluation, when comparing the performance of competitive surfers in order to reduce uncertainty and increase the fairness of these comparisons that can define the athlete's destiny in the competition. These results suggest the need to modify the competitive dynamics in the surf and, possibly, in other judging modalities.

Contributions

Contributed to conception and design: TMS
Contributed to acquisition of data: TMS, LERS, IV, CB, LCP, PHM, TCBM
Contributed to analysis and interpretation of data: TMS, LERS, IV, CB, LCP, PHM, TCBM
Drafted and/or revised the article: TMS, LERS, IV, CB, LCP, PHM, TCBM, HME, FOP
Approved the submitted version for publication: TMS, LERS, IV, CB, LCP, PHM, TCBM, HME, FOP

Acknowledgements

The authors of this study thanks to Vinicius de Oliveira Damasceno for his collaboration in the data extraction. We are also grateful for the considerations and reflections provided by Maicon

Rodrigues Albuquerque regarding strategy and the features of Rstudio in data analysis and Wagner Prado for overall comments. L.E.R.S., L.C.P., and T.C.B.M. are grateful for scholarship by CAPES Brazil. I.V. and C.B. are grateful for scholarship by FAPESP Brazil (#2021/07036-7 and #2020/04827-0).

Funding information

This research did not receive any type of funding.

Data and Supplementary Material Accessibility

The data, scripts and Supplementary Material used are available on https://osf.io/yk2vt/.

REFERENCES

[1] World Surf League. (2022). *Rules and regulations*. Available: https://www.worldsurfleague.com/pages/rules-and-regulations

[2] D. Kahneman and S. Frederick, "Representativeness revisited: Attribute substitution in intuitive judgment," in *Heuristics and biases: The psychology of intuitive judgment.*, ed New York, NY, US: Cambridge University Press, 2002, pp. 49-81.

[3] C. Staats, K. Capatosto, R. A. Wright, and D. Contractor, *State of the science: Implicit bias review 2015* vol. 3: Kirwan Institute for the Study of Race and Ethnicity Columbus, OH, 2015.

[4] S. Heiniger and H. Mercier. Judging the judges: A general framework for evaluating the performance of international sports judges [Online]. Available: https://arxiv.org/abs/1807.10055.

[5] K. Flessas, D. Mylonas, G. Panagiotaropoulou, D. Tsopani, A. Korda, C. Siettos, *et al.*, "Judging the judges' performance in rhythmic gymnastics," *Medicine and Science in Sports and Exercise,* vol. 47, pp. 640-8, Mar 2015. https://doi.org/10.1249/mss.0000000000000425.

[6] H. Hill and S. Windmann, "Examining Event-Related Potential (ERP) correlates of decision bias in recognition memory judgments," *PLoS One,* vol. 9, p. e106411, 2014. https://doi.org/10.1371/journal.pone.0106411.

[7] H. Plessner and T. Haar, "Sports performance judgments from a social cognitive perspective," *Journal Psychology of Sport Exercise,* vol. 7, pp. 555-575, 2006. https://doi.org/10.1016/j.psychsport.2006.03.007

[8] M. Muraven and R. F. Baumeister, "Self-regulation and depletion of limited resources: does self-control resemble a muscle?," *Psychol Bull,* vol. 126, pp. 247-59, Mar 2000. https://doi.org/10.1037/0033-2909.126.2.247.

[9] C. Leandro, L. Avila-Carvalho, E. Sierra-Palmeiro, and M. Bobo-Arce, "Judging in Rhythmic Gymnastics at Different Levels of Performance," *Journal of Human Kinetics* vol. 60, pp. 159-165, Dec 2017. https://doi.org/10.1515/hukin-2017-0099.

[10] M. B. Pajek, I. Cuk, J. Pajek, M. Kovač, and B. Leskošek, "Is the quality of judging in women artistic gymnastics equivalent at major competitions of different levels?," *Journal of Human Kinetics,* vol. 37, pp. 173-81, 2013. https://doi.org/10.2478/hukin-2013-0038.

[11] J. Premelč, G. Vučković, N. James, and B. Leskošek, "Reliability of Judging in DanceSport," *Front Psychol,* vol. 10, p. 1001, 2019. https://doi.org/10.3389/fpsyg.2019.01001.

[12] B. Leskošek, I. Čuk, I. Karácsony, J. Pajek, and M. Bučar, "Reliability and validity of judging in men's artistic gymnastics at the 2009 university games," *Science of Gymnastics Journal,* vol. 2, pp. 25-34, 2010.

[13] S. Heiniger and H. Mercier, "Judging the judges: evaluating the accuracy and national bias of international gymnastics judges," *Journal of Quantitative Analysis in Sports,* vol. 17, pp. 289-305, 2021. https://doi.org/10.1515/jqas-2019-0113.

[14] W. G. Hopkins, "Measures of reliability in sports medicine and science," *Sports Medicine,* vol. 30, pp. 1-15, Jul 2000. https://doi.org/10.2165/00007256-200030010-00001.

[15] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine,* vol. 15, pp. 155-63, Jun 2016. https://doi.org/10.1016/j.jcm.2016.02.012.

[16] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. [Online]. Available: https://www.R-project.org/

[17] W. Revelle. Procedures for personality and psychological research [Online]. Available: https://CRAN.R-project.org/package=psych

[18] A. G. Copay, B. R. Subach, S. D. Glassman, D. W. Polly, Jr., and T. C. Schuler, "Understanding the minimum clinically important difference: a review of concepts and methods," *Spine Journal,* vol. 7, pp. 541-6, Sep-Oct 2007. https://doi.org/10.1016/j.spinee.2007.01.008.

[19] G. Atkinson and A. M. Nevill, "Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine," *Sports Medicine,* vol. 26, pp. 217-38, Oct 1998. https://doi.org/10.2165/00007256-199826040-00002.

[20] T. H. Lyngstad, J. Härkönen, and L. T. S. Rønneberg, "Nationalistic bias in sport performance evaluations: An example from the ski jumping world cup," *European*

Journal for Sport and Society, vol. 17, pp. 250-264, 2020/07/02 2020. https://doi.org/10.1080/16138171.2020.1792628.